Titre de l'article : Force Concept Inventory : A Lesson to Be Learned
Titre de la recherche : Changes in Student Knowledge Structures in Science

Collège responsable : Vanier College

Chercheuse responsable : Helena Dedic

Notre numéro du projet : PA1996-1483
Date de réception à notre direction (MEQ – Enseignement supérieur) : 18 avril 2000.

Document disponible en ligne :
Centre de documentation collégiale
~~URL = http://www.cdc.qc.ca/textes/728774_dedic_article_parea_2000.doc~~
URL = http://www.cdc.qc.ca/pdf/728774_dedic_article_parea_2000.pdf

Force Concept Inventory: A Lesson to Be Learned

Helena Dedic, Miriam Cooper, Steven Rosenfield, Philip C. Abrami

Abstract: We investigated the effect of a short intervention on FCI performance. The intervention was a short training session designed to teach students to recognize FCI-type questions as ones that should be answered using physics knowledge. The experiment contrasted the performance of an experimental class with a control class. The results supported the hypothesis that student performance on the FCI is confounded by the format of its questions. Furthermore, we showed that the training session impacted low scoring students more than high scorers. Our experiment demonstrates that students should be familiar with the test format in order to obtain an accurate assessment of their understanding of Newtonian physics.

In the midst of the debate in the physics community as to what the Force Concept Inventory (FCI)(Hestenes, Wells, Swackhammer, 1992) actually measures, we want to report on our investigation of the effect of a training intervention on FCI scores. In our own research on conceptual change (as defined by Posner, et al., 1982; Strike & Posner, 1992), in physics, we attempted to develop an instrument to measure conceptual understanding. While interviewing students during the validation process, we became aware that test-construction issues such as item format and the precise wording of qualitative questions affected student interpretation of what was expected of them, and how they should answer. It was clear that these issues were interfering with our ability to measure their conceptual understanding with validity. Reflecting on our instrument's similarities to the FCI, we wondered whether these same issues might be confounding variables in students' scores on that test. In other words, we questioned whether the reported poor performance by students on the FCI (Hestenes et al., 1992, Hake, 1994) is solely due to students' incorrect conceptual understanding, or whether it is, at least in part, a reflection of the fact that the format of test items on this instrument is very different from that of standard physics class problems. If the latter were the case, we hypothesized that students' FCI performance would improve as a result of a short training session teaching them to recognize the novel question format as physics, and to apply their physics knowledge to solve such questions. Our report is on the results of an experiment testing this hypothesis.

Background

Hestenes et al. (1992) developed the FCI to assess student understanding of the many facets of the Newtonian concept of force. The FCI is a multiple-choice test which gives students a choice between one solution derived from a correct understanding of Newtonian mechanics and four other solutions based on "commonsense" alternatives (often referred to as misconceptions; a compiled list of papers on misconceptions in physics until 1998 can be found at http://www.oise.utoronto.ca/~science/physmisc.htm). In this paper we will use the term

"FCI-like" to refer to problems like those on the FCI while we use the term "physics-like" to refer to quantitative problems typically found in traditional physics texts and exams.

Although a problem on the FCI is of a qualitative nature and looks simple to a professional physicist, when the FCI was used to test students ranging from high school to university level across the US and Canada (Hake, 1994), it was found that students are less successful in choosing the correct qualitative answer than they would be in calculating the correct answer for a physics-like problem (Mazur, 1997). These results were considered to indicate that students, despite being successful in solving problems faced in traditional courses, retain their naive conceptions about the physical world (McDermott, 1991, 1993) and thus fail to correctly solve FCI-like problems. Because the FCI was accepted as a diagnostic tool of student understanding of the concept of force, the results caused dismay in the physics-teaching community, and promoted the use of interactive engagement (IE) methods in physics instruction intended to improve conceptual understanding.

Recently, concerns have been raised as to what the FCI actually measures (Huffman & Heller, 1995; Steinberg & Sabella, 1997), and what lessons one should draw from the findings (Griffith, 1997). The results of testing students using the FCI may be influenced by a number of factors such as its multiple choice format, the qualitative nature and wording of the problems, and whether the method of instruction is traditional or interactive engagement in nature.

Multiple choice format of the FCI

Steinberg and Sabella (1997) suggested that performance on the FCI may be influenced by its multiple-choice nature which may trigger responses students would not themselves generate.

Another issue came to light in our research. Students often solve multiple choice questions using an elimination strategy, where choices are only marginally influenced by any conceptual understanding. To illustrate the kind of thinking process involved, here is a transcript from an interview with an 'A' student who had recently graduated from Calculus-based Mechanics. The student was describing the way he answered FCI item #23 (see Appendix 1):

> Student "I'll be honest with you guys. If this was a test situation I know that these two are wrong, because I know that there's only one choice of this (points to **d**), one choice of this (points to **e**) and three choices of this **a**, **b** and **c**. Then it's more likely that these are the ones that are going to be right. Maybe you can't really follow."
> Researcher: "I follow".
> Student: "I suppose if somebody knows the answer to this quite well, he'll have no trouble knowing these are completely wrong, so he'll have no trouble erasing these (points to **d** and **e**) and he'll be narrowed down to these three answers (points to **a**, **b**, **c**). And he'll have no trouble with these three answers. What makes it even easier, is that it's kind of given away by this (points to path **e**). Once they've figured this kind of trick out, they would say 'of course, it goes in this kind of hyperbolic path'. If you were to give me this, this straight line, I would say this can't possibly be right. But since you have a hyperbolic kind of movement for this one (points to **e**), I would have chosen this one (points to **c**), so I know that this kind of path is right."

Although the student uses logical reasoning based on experience with multiple choice tests to choose the correct path, it is difficult to judge whether he has an understanding of the physics concept involved.

Qualitative nature of the FCI

Mazur (1997) raised the issue of the differences between qualitative and quantitative questions. He showed that if quantitative and qualitative problems on the same concept were paired, many students in a traditional lecture class demonstrated a serious conceptual misunderstanding in their solution to the qualitative version of the problem despite being able to produce the correct numerical solution in the quantitative version. He attributed this difference in their performance to students' using "recipes" or "algorithmic strategies" in solving physics-like problems without developing an underlying conceptual understanding.

Steinberg and Sabella (1997) studied students' performance on qualitative open-ended exam questions. They found that certain students performed better on these problems than they did on the FCI, even though they were matched for conceptual content and difficulty. They speculated that the wording of the FCI-like problems invoked thoughts of real world experiences, while the wording of their open-ended exam problems invoked thoughts of physics-like problems.

In another study[1] we interviewed students on their thinking when solving FCI-like problems versus physics-like problems. When students were given a qualitative and a quantitative version of the same problem, even if they had not received instruction in relevant concepts, they all readily answered the qualitative version while refusing to attempt to solve the quantitative version. When they had received instruction in relevant concepts, they tended to use different strategies, deducing the solution from their own experience to answer qualitative problems but using physics procedures to solve quantitative problems. For example, we gave a class of students two versions of a problem where a can is dropped from a moving car, one qualitative and one quantitative. For half the class, the qualitative version preceded the quantitative one while the order was reversed for the other half. Fragments from the transcript of a typical interview exemplify the different approach used by a student when answering, back-to-back, the same conceptual problem worded first qualitatively and then quantitatively.

> *FCI-like problem: A driver of a car travelling North at a steady 30m/s drops an empty Coke can. The diagrams below show the car at the moment the can is released. The dashed lines represent possible paths of the Coke can. Discuss the path in each diagram in terms of how likely you think the Coke can is to follow that particular path.*[2] *Explain your reasoning in each case.*

[1] The results obtained in that study will be published shortly.

[2] Note that suggested paths **a** ... **e** correspond to the paths in FCI item 23. Neither problem shows the path of the moving vehicle after the drop.

Student: "... If he dropped it in the car, then it would just drop to the bottom. There would be this one (points to **d**) .... But if he held his hand out the window and dropped it, then the car would continue to move forward and it would drop to the ground behind him and he would have passed it already (chooses **e**)." ...
Interviewer: "... you are eliminating these three (paths). Why?"
Student: "... you couldn't drop a can out the window and have it end up further "ahead" of you than the car is, and in all three of these, although they are different shapes, the can ends up in front of the person.

Physics-like problem: A passenger dropped an empty beer bottle from a train travelling at 40m/s headed due south. The bottle was dropped from a point 2 m above the ground. Determine the horizontal distance the beer bottle travelled before landing.

Student: "What's gonna happen to the bottle is that it starts out here and it's gonna end up going like that (*draws diagram depicting path similar to* **c** *in the previous qualitative question*), down to the ground. And this is gonna be 2 m, and we are looking for this horizontal distance. ... It started with a horizontal velocity of 40 m/s ..." (*The student then goes on to solve the problem. In his discourse he displayed a firm grasp of the concept of inertia.*)

Note the different thought process indicated by the student's responses to the two questions. He responds to the FCI-like problem by recalling his own experience of being in a car, which influences his interpretation of the picture since he argues against the can going "ahead" of the car. For the physics-like problem, however, he sketches without hesitation a trajectory that is identical to path **c** of the FCI-like problem, and then proceeds to draw on his knowledge of physics to formulate the correct answer. The student demonstrated no awareness of the blatant contradiction between his responses, and expressed puzzlement when the inconsistency in his thinking was pointed out by the interviewer. It appears that cues such as "determine" and "the horizontal distance" elicited a link to problem-solving strategies in his approach to the physics-like problem. Instruction that teaches students to attend to physics cues, rather than to personal experiences, might have helped this student.

Interactive engagement methods

Hake (1998, p.65) identified "IE methods as those designed at least in part to promote conceptual understanding through interactive engagement of students in heads-on (always) and hands-on (usually) activities which yield immediate feedback through discussion with peers and/or instructors". Traditional courses are those that "make little or no use of IE methods, relying primarily on passive-student lectures, recipe labs and algorithmic-problem exams". Hake's survey of 1998 indicated that performance gains on the FCI were higher for students enrolled in courses which made substantial use of IE methods. Numerous studies of IE methods (e.g., peer instruction (Mazur (1997), workshop physics (Laws, 1997)) have shown them to be effective in improving students' performance on qualitative questions and the FCI.

The above results imply that IE instruction is better than traditional instruction at promoting conceptual understanding of Newtonian concepts as measured by the FCI. We wonder if an alternative explanation might be that the better performance on the

FCI by students in IE courses is due, at least in part, to their exposure to qualitative questions within that instructional setting. Students enrolled in traditional lecture courses may have less experience in answering qualitative problems than students in IE courses. Although most physics text books include qualitative problems called "Questions" that precede the standard numerical problems, many teachers do not assign such questions or use them on tests (Dicke, 1997). It is likely that FCI-like problems are discussed in class and posed on exams in IE courses. Browsing through IE textbooks (McDermott (1996), Knight (1997), Mazur(1997)), we found support for this conjecture.

We wanted to clarify to what extent the aforementioned issues might play a role in the FCI scores and decided to test whether training students to use appropriate problem-solving strategies when answering FCI-like questions would improve their score on the FCI. The training would not be designed to teach concepts, but rather to demonstrate to students that they should use the same strategies to answer FCI-like questions as they do to answer physics-like problems.

The Experiment

Participants

We compared the FCI performance gains of two groups of students taught in college-level introductory physics by the same instructor. Both classes followed the same curriculum and used the traditional lecture format in class. The students had comparable academic profiles with high school science averages between 65% and 70%, which is why they were required to enroll in a remedial program which includes a pre-calculus course, an introductory chemistry course, an introductory physics course (Introduction to College Physics) and a learning-to-learn course (Introduction to College Science). The experimental class of 34 students ran in the fall term of 1997, while the control class of 22 students ran in the fall term of 1998.

Design

In this quasi-experimental study we used a 2X2 mixed factorial design with one between-group factor with two levels (experimental, control) and one within-group factor (pre-test, post-test).

Pre-test/Post-test:

The researchers administered the FCI as a pre-test (Pre) during a regular lab period two weeks before the end of the course in both groups. All students were given the FCI as a post-test (Post) immediately following their final exam.

During an interview the physics instructor revealed that he did not teach all the material pertinent to all FCI items. Therefore, we divided the 29-item FCI into two subsets: Set A includes all questions (19 items) relevant to material taught in the course, while Set B includes all other questions (10 items). In the results below we examine student performance on the entire FCI questionnaire, as well as on the two subsets.

Training

A 75-minute training session on answering FCI-like questions was given to the experimental class one week after the pre-test. The session was given by a member of our research team and used IE methods. It should be noted that the researcher had used IE methods all semester with the same students as their teacher of Introduction to College Science. Even though the session was only 75 minutes long, the trust that existed between the instructor and the students made them receptive to her intervention. Instead of a training session, the control class received a regular 75 minute tutorial in preparation for the final exam from their physics instructor.

The instructional materials consisted of a set of nineteen problems. We generated the problem set by translating physics-like problems from the course text into FCI-like problems. All problems were presented in the same format as those on the FCI, but were different from FCI items in order to avoid "teaching to the test". We examined student performance on each item to assess whether the effect of training was due more to the content of the practice problems than due to strategies used. The nineteen practice problems were chosen before we were aware which FCI material was covered in the course and which was not.

During the session the researcher used IE methods and started by modelling the solution of one problem. The researcher emphasized that FCI-like problems are solved by first drawing a diagram of the situation and then thinking of the physics involved, not by choosing an answer that made sense from their own perceptions or by using elimination strategies. The researcher pointed out that the problems described real-world situations using colloquial language which may jog memories of real-world experiences. The class discussion contrasted the unreliability of individual perceptions to the predictive power of Newtonian principles. In conclusion, the researcher told the students to wear their physicists' hats and use the same strategies for solving FCI-like problems as physics-like problems in order to be successful on the FCI. The second problem was done in discussion with the entire class. The third problem was done by small groups reporting to the whole class. After the instruction students were encouraged to finish the remaining eighteen problems at home, and told that they could pick up solutions or seek help if needed from the researchers. A number of students picked up the solutions, but only a few sought additional help. A sample problem follows:

body is fired upward with initial velocity $\mathbf{v}_0$. It takes time T to reach its maximum height H. Which statement is true?

a. It takes half the time (T/2) to reach half its maximum height (H/2.
b. It takes half the time (T/2) to decrease to half its initial speed (v_0/2).
c. It has half the initial speed (v_0/2) when it reaches half its maximum height (H/2).
d. It has the same velocity just before it lands as when it was fired.

Practice Problem

Student incentive

Steinberg and Sabella (1997) felt that the difference they observed in student performance between the FCI and final exam might be due in part to the fact that the FCI did not count towards students' grades. Consequently, as an incentive to do their best on both tests, students in our study were told that they could earn up to 5 bonus points towards their final grades as a function of their FCI scores.

Results

Equivalence of experimental and control classes

The experimental and control groups can be considered equivalent if it can be shown that there was no difference between the two groups' performance on the FCI pre-test. Table 1 shows the average pre-test score <Pre_i> and the standard error of the mean (*sem*) for the experimental and control groups. The mean FCI pre-test score was 41.68% (sem 2.49) for the experimental class, and 41.69% (sem 3.19) for the control class. A two-tailed t-test for independent samples yielded a probability, P_t, of 0.998 that there is no significant difference between these two means.

Table 1
Comparison of the Pre-test Scores for experimental and control classes

Measure	Experimental	Control	P_t
<Pre_i> (*sem*)	41.68% (*2.49*)	41.69% (*3.19*)	0.998

Improvement in FCI score.

We examined the change between the mean pre-test score <Pre_i> and the mean post-test score <Post_i> for both the experimental and control conditions. We found that there was significant improvement in both classes between the pre-test and the post-test score. The mean FCI score rose from 41.68% to 52.64% for the experimental class, and from 41.69% to 46.40% for the control class. The one-tailed t-test for repeated measures showed that the change was significant for both classes (t(33) = 5.63, p<0.001, t(21) = 3.70, p<0.001). Similarly, the mean FCI score increased significantly.

In order to see whether students performed differently on subsets of items covered (Set A) and not covered (Set B) in the course, we examined the change between the mean pre-test score and the mean post-test score for the two subsets. The mean score for Set A rose significantly from 42.26% to 57.12% for the experimental class (t(33) = 4.96, p<0.001), while the mean Set A score increased from 45.7% to 49.28% for the control class (t(21) = 1.65, p<0.1). We noted that the improvement in scores on Set B was not statistically significant (t(33) = 1.16, p<0.1) for the experimental class. On the other hand, the improvement in scores on Set B (the mean pretest score is 34.09 and the mean post-test score is 40.91) is significant (t(21) = 2.05, p<0.05) for the control class. Table 2 shows these results.

Table 2
Comparison of mean FCI scores for Experimental and Control Classes

	Exp. FCI	Control FCI	Exp. Set A	Control Set A	Exp. Set B	Control Set B
<Pre$_i$> *(sem)*	41.68% *(2.49)*	41.69% *(3.19)*	42.26% *(2.82)*	45.7% *(3.79)*	40.59% *(2.85)*	34.09% *(3.56)*
<Post$_i$> *(sem)*	52.64% *(2.5)*	46.40% *(3.7)*	57.12% *(2.57)*	49.28% *(4.43)*	44.12% *(2.57)*	40.91% *(4.21)*

Effect of training on FCI score

The difference between the mean gains <G$_i$> = <Post$_i$ - Pre$_i$> for the experimental class and the control class is a measure of the effect of training. However, gains may also be affected by the ceiling effect and, consequently, may decrease as pre-test scores increase. To account for this possibility and control for it statistically we also computed the normalized gain per student as the ratio of the actual gain G$_i$ =Post$_i$ -Pre$_i$ to the maximum possible gain $G_{i\,\max}$ =100-Pre$_i$. The mean normalized gain, <g$_i$>, is

$$<g_i>=<G_i/G_{i\,\max}>=<(Post_i - Pre_i)/(100 - Pre_i)>.$$

The coefficient of correlation *r* was also calculated to measure the correlation between gains G$_i$ and pre-test scores Pre$_i$. This assesses how gains vary with pre-test scores and whether the relationship between the gains and the pre-test scores changed as a result of the training.

Gains: Table 3 shows the mean gains and the normalized mean gains on the whole FCI, as well as on the subset of items pertaining to material covered in class (Set A) and on items not covered in class (Set B) in both the training and control conditions. We found that the mean gain <G$_i$> (10.95%) in the experimental class is significantly higher than the mean gain (4.70%) in the control class (one-tailed t-test for independent samples: t(54) = 2.38, $p<0.05$). When we examined the difference between the mean gain (14.86%) for the experimental class and the mean gain (3.58%) for the control class, we again found that the gains were significantly higher in the experimental (t(54) = 2.73, $p<0.01$) for Set A while the difference was not significant (t(54) = 0.71, $p<0.1$) for Set B (the mean gain is 3.53% in the experimental and 6.82% for the control class).

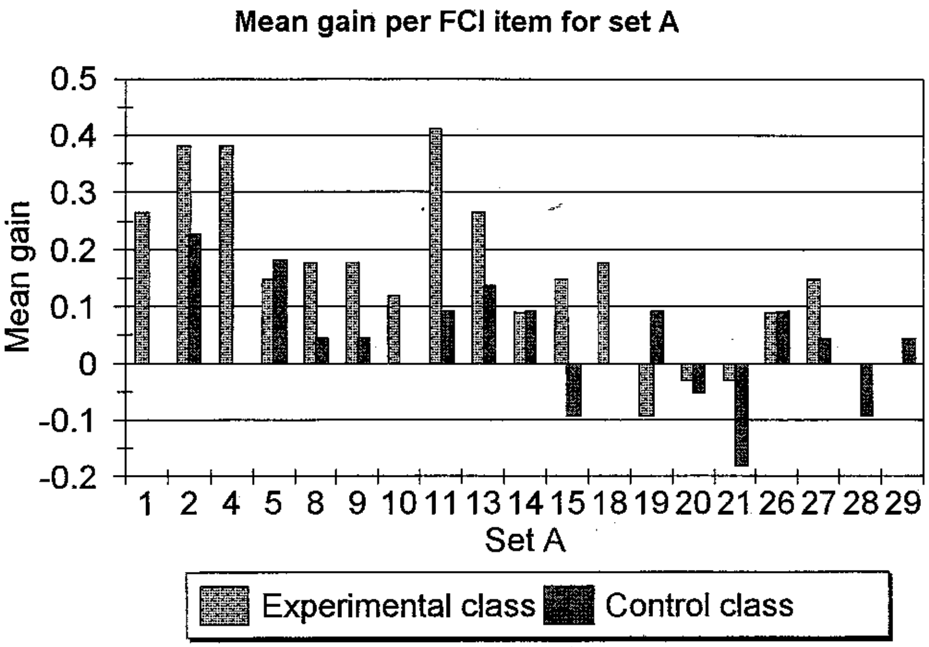
Normalized gains: The examination of the normalized gains yielded similar results. We found that the mean normalized gain <g$_i$> (0.19) in the experimental class is significantly higher than the mean normalized gain <g$_i$> (0.1) in the control class (a one-tailed t-test for independent samples: t(54) = 1.82, $p<0.05$). When we examined the difference between the mean normalized gain (0.24) for the experimental class and the mean normalized gain (0.06) for the control class on Set A, we found that the normalized gains were also significantly higher (t(54) = 2.69, p<0.01) for the experimental class while the difference was not significant (t(54) = 0.47, p<0.1) on the questions of Set B (the mean normalized gain is 0.06 for the experimental and the mean normalized gain is 0.01 for the control class).

Table 3
Comparison of mean gains for Experimental and Control Classes

	Exp. FCI	Control FCI	Exp. Set A	Control Set A	Exp. Set B	Control Set B
$<G_i>$ *(sem)*	10.95% *(1.91)*	4.70% *(1.25)*	14.86% *(2.95)*	3.58% *(2.12)*	3.53% *(2.30)*	6.82% *(3.24)*
$<g_i>$ *(sem)*	0.19 *(0.03)*	0.10 *(0.03)*	0.24 *(0.04)*	0.06 *(0.05)*	0.06 *(0.10)*	0.10 *(0.06)*

The impact of training on students' knowledge: We wanted to be certain that any gains made by experimental students were not due to knowledge acquired during the training session. To this end, we examined the mean gain per FCI item in both classes (see Graph 1).

Graph 1

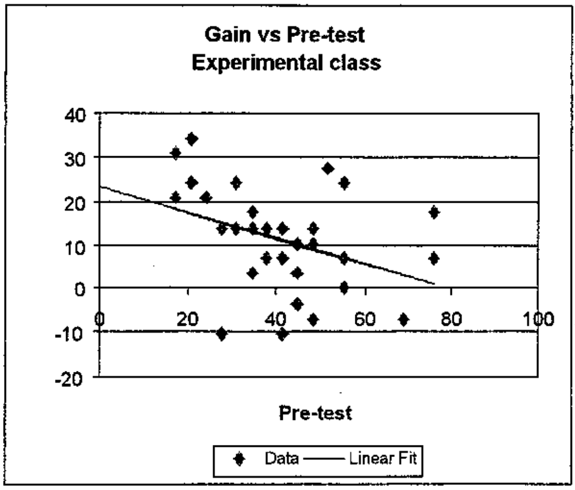


We found that the experimental students made noticeably greater gains than the control students on three particular items (#1, #4 and #11). We carefully examined the practice problems to see whether students could have acquired particular knowledge that would account for such a difference in the performance. There was no problem in the practice set similar to FCI item #1. There was a problem on topics related to each of items #4 and #11, although these were not covered in the training session and did not call for the students to answer the same questions as on the FCI. Nonetheless, we decided to rerun all the statistical tests after having removed items #4 and #11 from the data. All relevant statistics were still significant, indicating that content did not play a role in the training effect.

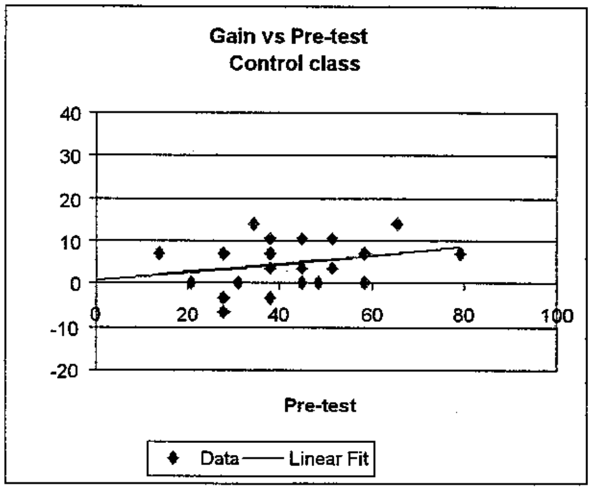
Relationship between the gain and the pre-test: We found that in the experimental class, gains, G_i, decreased with pre-test scores, Pre_i, with a correlation coefficient of $r = -0.377$. This relationship was reversed in the control class, where the gains increased with pre-test scores ($r = 0.249$). To assess the significance of this result, we used Cohen's (1998) conventions and determined the power of significance[3] for both classes. We found that the power of significance is 55%, given the effect size and the number of subjects in the experimental class, and similarly 20%, given the effect size and the number of subjects in the control class.

We also examined the correlations between pretest scores and gains, G_i for items in Set A and in Set B. There was a marked difference in the correlation between gains and pre-test scores for Set A between the two classes: in the experimental group $r = -0.694$ (power of significance 97%) and in the control class $r = 0.48$ (power of significance 65%). For Set B items there was no correlation in the experimental class ($r = -0.020$) and a low correlation in the control class ($r = -0.235$, power of significance 20%). For Set A, Graphs 2 and 3 are plots of G_i versus Pre_i for the experimental class and control class, respectively.

Graph 2



Graph 3



[3] Power of significance is a measure of significance which depends on the effect size and the sample size. When the effect size is large, the result may be significant even for a small sample size.

Effect of training in context of other studies:

Since both the pre-test and post-test were administered at the end of the course with only two weeks between them, we cannot compare our results to the usual pre/post-test data where the tests are administered at the beginning and end of the course. Even so, it is interesting to see our results with respect to those from the usual pre/post-test condition.

In his survey of physics courses, Hake (1998) used <g> rather than $<g_i>$ as a measure of the average normalized gain on the FCI, where <g> is the ratio of the actual average gain to the maximum possible average gain.

$$<g> = <G_i>/<G_{i\,max}> = (<Post_i> - <Pre_i>)/(100 - <Pre_i>)$$

As Hake points out, somewhat lower random errors are entailed if one takes the average normalized gain for a course to be $<g_i>$ rather than <g>.

In Table 4, we show <g> for our experimental and control classes with associated error Δ<g> which was computed using the same formula as the one used by Hake. For comparison we show $<<g>>_{T\text{-}ave}$ and $<<g>>_{IE\text{-}ave}$, the average values of <g> for fourteen traditional classes and forty eight IE courses taken from the Hake survey.

Table 4
Comparison with Hake's Data

	Experimental <g> ± Δ<g>	Control <g> ± Δ<g>	Hake survey $<<g>>_{T\text{-}ave}$ ± sd	Hake survey $<<g>>_{IE\text{-}ave}$ ± sd
FCI	0.19 ± 0.06	0.08 ± 0.08	0.22 ± 0.05	0.52 ± 0.10
Set A	0.26 ± 0.07	0.07 ± 0.11		
Set B	0.06 ± 0.08	0.10 ± 0.09		

Discussion

Equivalence of experimental and control classes

Students were not randomly assigned to experimental and control classes. However, there are indicators that the two classes were statistically equivalent. The high-school academic profiles of students in both classes satisfied the same narrow admission criteria (high school average between 65% and 70%) placing them in a remedial program. However, since students come from a variety of schools there is a possibility of a larger spread in their academic performance than is shown by their high school grades. The pre-test instrument (FCI) was the same for both classes and was administered under the same conditions. Consequently, the fact that the mean pre-test scores were statistically equal for the two classes is a strong objective indicator that the two classes were indeed equivalent. The pre-test scores were low in both classes (41.68% and 41.69 % for the experimental and control class respectively) as might be expected from students in a remedial program.

Improvement in the FCI score.

Since the pre-test and post-test were administered within the last two weeks of classes, we expect a gain on the post-test to be due to both students' preparation for the final physics test and to pre-test exposure to the FCI. Indeed, both classes had significant gains.

Effect of training on the FCI score

Both the gain and the normalized gain are significantly larger for the experimental class than for the control class. The training was effective in improving student performance.

We found that the mean gains per item were similarly distributed (Graph 1) with the mean gain decreasing with item number (which we suspect was due to fatigue since the students wrote the post-test immediately following the final exam). We conclude that it is unlikely that the training had an effect on the domain of knowledge. We therefore believe that the impact of the training was primarily on strategies students used to answer the questions.

This belief is further supported by the results for the two subsets. For Set A, the subset of items relevant to material taught in the course, there was a larger gap between the mean normalized gain for the experimental class and the mean normalized gain for the control class as compared to the gap between those means for the whole set of the FCI. For Set B, the subset of items relevant to material not taught in the course, there was no significant difference between the gains and the normalized gains for the two classes. This indicates that the training was only effective when the students had appropriate knowledge.

We also found a significant effect of the training on the correlation between gains and pretest scores. We will limit our discussion to the results for Set A where the correlation coefficients are large and significant and where we may be more confident that students have the conceptual knowledge of Newtonian concepts. There was a strong negative correlation between gain and pre-test score for the experimental class and a positive correlation for the control class.

If we consider the relationship between gains and pre-test scores, we anticipate three possible factors at work: the ceiling effect; student preparation and the training effect. In the analysis of our data we were able to discount the ceiling effect since only one student reached the ceiling and consequently, we don't expect that this factor plays an important role in the relationship between the gains and the pre-test scores. The second factor is the effect of student effort and preparation for the final test. If this factor were to play a role, we would expect gains to correlate positively with pre-test scores. The good students have a tendency to work harder and learn more in preparation for finals than poor students. Consequently, high scorers on the pre-test are likely to have higher gains. The third factor is the effect of training. Low pre-test scorers lack conceptual understanding, or strategic knowledge or both. The training should have an effect on those who only lack strategic knowledge and thus, fall into the trap of not using Newtonian concepts to answer FCI-like problems. Since the high pre-test scorers are likely to have both conceptual understanding and strategic knowledge, we expected that the training would be most effective for students scoring at the low end of the pre-test score range. If this factor were to play a role, we would expect gains to correlate negatively with pre-test scores.

The results show that the anticipated effect of student preparation was evident in the positive correlation between gains and pre-test scores in the control class. In the experimental class, however, we see a strong negative correlation between the two variables. While we still imagine that there was an effect of student preparation on the correlation, it was outweighed by the impact of training. This indicates that there were students in the experimental class who lacked strategic knowledge and who improved their scores to reflect their true conceptual understanding. There was not much difference in gains for students who scored above 50% on the pre-test between the two classes. This indicates that the training had less impact on high scorers.

Effect of Training in Context of other studies

It is noteworthy that for the experimental class, the average normalized gain, where the gain was made over two weeks, compares favourably to the average for traditional courses (Hake, 1998) where the gain was made over the entire period of the course. The gain does not compare to the gains made in the IE classes, which is not surprising since our IE intervention lasted only 75 minutes.

Conclusions

Our experiment shows that there is a confounding factor in using the FCI as a measure of conceptual understanding. The results of this study are of particular importance to instructors in traditional lecture based classes who wish to use the FCI to assess their students understanding of Newtonian physics. It indicates that the performance of their students on the FCI is not necessarily an accurate reflection of their conceptual understanding. It is important that students be tested with questions whose format is familiar to them. We have shown that a short training session can provide students with appropriate strategies so that the FCI may more accurately reflect conceptual understanding.

Although we ourselves have used IE methods extensively in our classes for 20 years, traditional methods are standard in the physics department. We were struck by the fact that even though the instructor of the classes in this study used traditional methods, his students showed their conceptual knowledge when they had the appropriate strategic knowledge. In particular, the instructor stressed Newton's Third Law, and the students performed relatively well on FCI items #2, #13 and # 14. We are inclined to agree with Griffiths (1997) who does not believe that "traditional methods are hopelessly flawed".

We speculate that one reason that FCI mean gains in IE courses are higher than in traditional classes is that students are taught strategies for FCI-like problems. It would be interesting to know if the higher mean gains in IE courses are due to gains by all students or are mostly due to gains made by low scorers. If the latter is the case, then it may not be that IE courses are better than traditional courses in promoting conceptual understanding, but rather in providing skills for solving FCI-like problems. The question of whether traditional or IE courses are more effective pedagogies may not be resolved purely on the basis of FCI results.

BIBLIOGRAPHY

Dicke, L. O. (1997) *Assessment and Learning in Physics* . A paper presented at the annual meeting of the Association pout la Recherche au Collegial, Dawson College, Montreal, Quebec.

Griffiths, D. (1997) Millikan Lecture 1997: Is there a text in this class? *American Journal of Physics, 65,* 1043-1055.

Hake, R.R., (1994). Survey of test data for introductory mechanics courses. *AAPT Announcer 24,* 55.

Hake, R.R. (1998) Interactive-engagement vs traditional methods: A six-thousand-students survey of mechanics test data for introductory physics courses, *American Journal of Physics, 1,* 64-74.

Hestenes,D., Wells, M., & Swackhammer, G. (1992). Force Concept Inventory. *The Physics Teacher, 30,* 141-153.

Huffman, D., Heller, P. (1995). What Does the Force Concept Inventory Actually Measure? *The Physics Teacher, 33,* 142-143.

http://www.oise.utoronto.ca/~science/physmisc.htm

Knight, R. D. (1997). Physics, A Contemporary Perspective. Addison Wesley Longman, Inc.

Laws, P. W. (1997). *Workshop Physics Activity Guide , Modules 1-4 w/ Appendices,* John Wiley & Sons, New York, 855 pp.

Mazur, E. (1996). *Understanding of Memorization: Are we teaching the right thing.* Proc. Resnick Conference, in press, Wiley, 1996.

Mazur, E. (1997). Peer Instruction. Upper Saddle River NJ: Prentice Hall, Inc.

McDermott, L. C. (1996). *Physics by Inquiry Volumes I and II* . John Wiley & Sons, Inc., New York, NY.

McDermott, L. C. (1991) Millikan Lecture 1990: What we teach and what is learned-closing the gap. *American Journal of Physics, 59,* 301-315.

McDermott, L. C. (1993) Guest Comment: How we teach and how students learn- A mismatch:. *American Journal of Physics, 61,* 295-298.

Posner, G., Strike, K., Hewson, P., & Gertzog, W. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education, 66,* 211-227.

Steinberg, R. N., Sabella, M. S. (1997). Performance on Multiple-Choice Diagnostics and Complementary Exam Problems. *The Physics Teacher, 35,* 150-155.

Strike, K. A., & Posner, G. J. (1992). A revisionist theory of conceptual change. In R. Duschl & R. Hamilton (Eds.), *Philosophy of science, cognitive psychology, and* *educational theory and practice* (pp. 147-176). Albany, NY: State University of New York.